AN EFFICIENT ALGORITHM FOR DNA DISCRETE FOURIER ANALYSIS

Ahmad Rushdi and Jamal Tuqan

Department of Electrical and Computer Engineering
University of California, Davis
aarushdi@ece.ucdavis.edu, tuqan@ece.ucdavis.edu

ABSTRACT

A number of signal processing techniques have been recently proposed to locate periodicities in DNA sequences. A key building block in any of these methods is the computation of the M-point short-time discrete Fourier transform $X(n,k)$ at a particular frequency component $k = M/R$ where R is an integer. In this paper and using multirate signal processing theory, we present a computationally efficient approach to find $X(n, M/R)$. In specific, we first derive a new multirate DSP model that i) computes $X(n,R)$ for *any* given integer R, and ii) is completely parameterized by a set of real digital filters $H_r(z)$. We then obtain a closed-form expression for $X(n, M/R)$ that clearly quantifies the computational efficiency of the new method, and finally discuss the design of $H_r(z)$ to enhance the computation accuracy of $X(n, M/R)$.

1. INTRODUCTION

The application of signal processing techniques to analyze the structure of DNA sequences has been an active area of research in recent years. A fundamental component of such methods has been the deployment of the discrete Fourier transform (DFT) to identify periodicities within DNA sequences [1, 2, 3]. These periodicities are typically used to characterize important biological features. For example, the use of the magnitude squared of the DFT component at the frequency $\omega = 2\pi/3$ has been proposed to discriminate between protein coding and non-coding regions in a DNA sequence [4, 5, 6, 7]. A measure that depends on the phase of the same DFT component has also been suggested for gene finding [8, 9]. Other DNA periodicities were found in [10] and in [11] using the DFT and the warped DFT (W-DFT) respectively. Most of these methods however suffer from a high computational cost especially when searching over large DNA sequences in the order of thousands of nucleotides. The development of efficient computational tools for analyzing DNA sequences is a major requirement in genomic research. In this paper, we present a novel efficient algorithm that finds the short time DFT (ST-DFT) at *any frequency* $2\pi/R$ where R is an integer. The algorithm is based on a new DSP model that i) generalizes previous work (namely [6, 7]), ii) provides a closed form expression for the ST-DFT at $\omega = 2\pi/R$, and iii) reduces substantially the computational cost of this frequency component. The new scheme is also completely characterized by a set of real digital filters which, in turn, can be designed to enhance the accuracy of the frequency component. Although we mainly focus on DNA sequence analysis, our findings apply to many other DSP applications.

2. SYMBOLIC TO NUMERIC MAPPING

A single DNA strand is represented as a sequence of letters that belong to the alphabet $\mathbb{F} = \{A,\ C,\ G,\ T\}$. The letters are the standard symbols given to the four nucleotides: Adenine A, *Guanine* G, *Thymine* T, and *Cytosine* C. Along the two strands of the DNA double helix, a pyrimidine in one chain always faces a purine in the other with a specific pairing of the bases: T with A and, C with G. For example, a typical section of a DNA double-strand is shown in Figure 1, where the 5' and 3' are labels for the two ends of the nucleotide. Mapping a DNA sequence to a set of digital

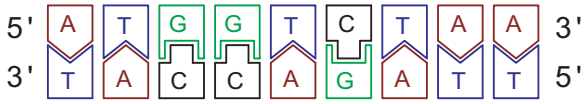

Figure 1: DNA double helix structure

signals is in general a common problem in the context of DNA data generation and several apparently distinct and unrelated approaches exist. The simplest mapping solution is the Voss representation which forms four binary indicator sequences $x_l(n), \forall\ l \in \mathbb{F}$ where a 1 would indicate the presence of a base and 0 indicates its absence. For example, the mapping of the top single DNA strand in Figure 1 into the indicator sequence $x_A(n)$ generates $\{1,0,0,0,0,0,0,1,1\}$.

3. A GENERAL MULTIRATE DSP MODEL

Assume that a discrete time sequence $x_l(n)$ has length N. The M-point ST-DFT of $x_l(n)$ is

$$X_l(n,k) \triangleq \sum_{m=0}^{M-1} x_l(n+m)e^{-j2\pi mk/M} \qquad (1)$$

where $n = 0, P, ..., (N-M+1)$ and P is the amount of window shift. Given the four binary indicator sequences $x_l(n), \forall\, l \in \mathbb{F}$, we can therefore find the corresponding DFT sequences $X_l(n,k), \forall\, l \in \mathbb{F}$. To determine the value of the spectrum at a specific frequency $\omega = 2\pi/R$ where R is an integer, we evaluate $X_l(n,k)$ at the index $k = M/R$. We assume here that the ST-DFT length M is a multiple of R, i.e. $M = qR$ for some integer q. From (1), it follows that

$$X_l(n) \triangleq X_l(n, \frac{M}{R}) = \sum_{m=0}^{M-1} x_l(n+m)e^{-j2\pi m/R} \qquad (2)$$

The direct evaluation of (2) is expensive since it involves the multiplication of the sequence samples by the M complex exponentials $\{e^{-j2\pi m/R}\}$. Furthermore, these operations are repeated with every window shift P. In the remainder of this paper, we show that such a computational cost can be in fact *greatly reduced*. To do this, let $P = R$ and note that equation (2) can be interpreted as the convolution of $x_l(n)$ with the complex digital filter $f(n)$

$$f(n) = \begin{cases} e^{-j\frac{2\pi}{R}n} & 0 \le n \le M-1 \\ 0 & otherwise \end{cases}$$

followed by decimation by R. Let $F(z) = \mathcal{Z}\{f(n)\}$ denote the $\mathcal{Z}$ transform of $f(n)$. We can then show that $F(z)$ can be expressed as *a two-stage filter* obtained by cascading the complex filter

$$C(z) = 1 + e^{j\frac{2\pi}{R}}z^{-1} + \cdots + e^{j(R-1)\frac{2\pi}{R}}z^{-(R-1)}$$

with the real interpolated filter

$$H(z^R) = 1 + z^{-R} + ... + z^{-(M-1)}$$

as illustrated in Figure 2 where the box labelled $\downarrow R$ is a downsampler with decimation ratio R. A number of remarks are in order at this point. First, let $M = \prod_{i=1}^{B} M_i$, where B is the number of M's prime factors. Then, the suggested model can be used to locate the $\frac{2\pi}{R}$-frequency component *iff* $R = \prod_{i=1}^{B} M_i^{n_i}$, where $n_i \in \{0,1\}$, $\prod_{i=1}^{B} n_i \neq 0$, and $M_i \neq 1, \forall\, i \in \{1,2,\ldots,B\}$. Second, we note that the complex filter $C(z)$ has $R-1$ zeros on the unit circle, starting at $\omega = 0$ and spaced by $\frac{2\pi}{R}$ except at $\omega = \frac{2\pi}{R}$ as shown in Figure 5(b). The *interpolated* version of $H(z)$, $H(z^R)$, produces frequency images at $\omega = 0,\ \frac{2\pi}{R}\ \ldots\ 2\pi\frac{R-1}{R}$ as given in Figure 6(f). By cascading the two filters, we can see from the above figures that $C(z)$ attenuates the images at the unwanted frequencies of $H(z^R)$ and leaves only one opening at the required frequency $\omega = \frac{2\pi}{R}$. The resulting filter $F(z)$ is therefore a *complex* band pass filter with center frequency $\omega = 2\pi/R$. Finally, by invoking the polyphase decomposition, we can derive the new model of Figure 3 where, in this case, $H_r(z) = H(z)$ for all $r = 0, 1, \ldots, R-1$. In the more general setting illustrated in Figure 3, the sequences $X_{lr}(n)$ is given by

$$X_{lr}(n) = h_r(n) * x_{lr}(n) \qquad (3)$$

and is termed the r^{th} *filtered polyphase component* of $X_l(n)$, defined $\forall\ l \in \mathbb{F}$, and $\forall\ r \in \{0,1,...,R-1\}$. From Figure 3, we can immediately observe that the new structure separates the real part of $F(z)$, namely $H(z)$, from its complex part, namely $C(z)$. More importantly, the complex part $C(z)$ is *fixed* and only $R \ll M$ complex computations are required in this case. Finally, we emphasize that the above scheme is a generalization of the DSP model derived in [7] for the specific case of $R = 3$. From Figure 3, we obtain closed form expressions for the filtered polyphase components $X_{lr}(n)$ and subsequently for the DFT $X_l(n)$ in terms of the polyphase components of $x_l(n)$.

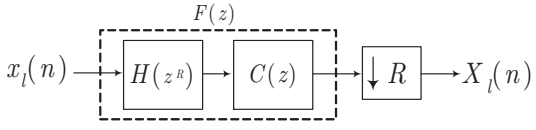

Figure 2: Cascade implementation of $F(z)$

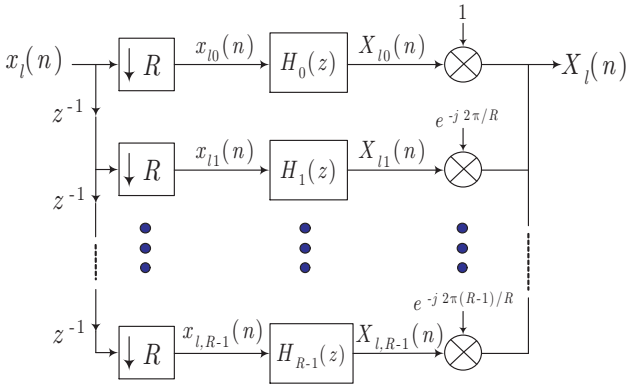

Figure 3: Multirate DSP model for $\frac{2\pi}{R}$-frequency component extraction

4. MATHEMATICAL ANALYSIS

We first split the discrete time sequence $x_l(n)$ into its R *polyphase components* with decimation ratio R to get $x_l(n+Rm)$, $x_l(n+Rm+1)$ ..., $x_l(n+Rm+R-1)$. These R polyphase components are generated by passing the signal $x_l(n)$ through an R-branch multirate blocking structure composed of a delay chain followed by downsamplers ($\downarrow R$) as shown in Figure 3. The polyphase sequences of the windowed sequence $x_A(n)$ with $R=5$ and $M=10$ are given in Figure 4.

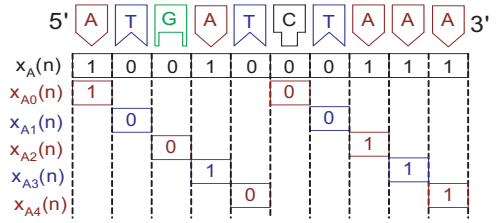

Figure 4: Polyphase sequences of a 10-sample window of $x_a(n)$ with $R=5$

We now can re-express $X_l(n)$ in (2) as follows

$$\begin{aligned} X_l(n) &= \sum_{m=0}^{\lfloor \frac{M}{R}-1 \rfloor} x_l(n+Rm) \\ &+ \sum_{m=0}^{\lfloor \frac{M}{R}-1 \rfloor} x_l(n+Rm+1)e^{-j\frac{2\pi}{R}} + \ldots \\ &+ \sum_{m=0}^{\lfloor \frac{M}{R}-1 \rfloor} x_l(n+Rm+R-1)e^{-j(R-1)\frac{2\pi}{R}} \\ &= \sum_{r=0}^{R-1} \sum_{m=r,r+R,\ldots}^{\lfloor \frac{M}{R}-1 \rfloor} x_l(n+Rm+r)e^{-jr\frac{2\pi}{R}} \quad (4) \end{aligned}$$

The inner summation in (4) represents $X_{lr}(n)$ and is simply the sum of the samples in the r^{th} codon position of $x_l(n)$ or, in other words, the number of occurrences of nucleotide l in codon position r of the window at index n. We can therefore define

$$X_{lr}(n) = \sum_{m=r,r+R,\ldots}^{\lfloor \frac{M}{R}-1 \rfloor} x(n+Rm+r), \quad (5)$$

and obtain the following closed form expression

$$X_l(n) \triangleq \sum_{r=0}^{R-1} X_{lr}(n)e^{-j2\pi r/R} \quad (6)$$

Note that equation (5) is indeed the convolution operation given in (3) where, in this case, $h_r(n)$ is a rectangular window. From (6), $X_l(n)$ is obtained by multiplying $X_{lr}(n)$ with the corresponding complex coefficients and adding the outputs as in Figure 3.

5. COMPLEXITY COMPARISON

To quantify the comparison between the direct DFT expression and the multirate model alternative, we compare the computation complexity in both cases. Equation (2) finds the frequency component at $\frac{2\pi}{R}$ using $\{M\}$ *complex* multiplications and $\{M-1\}$ *complex* additions for every nucleotide base l. Typical DNA sequences can be hundreds of thousands of bases which indicates that the direct computation of (2) is not a good alternative when analyzing huge databases. On the other hand, equation (5) can be evaluated using $R\lfloor \frac{M}{R} \rfloor \simeq M$ real additions. Note that these additions are actually simple increments since we are adding up the elements of a binary sequence. In turn, equation (6) requires $\{R\}$ complex multiplications and $\{R\}$ complex additions. Therefore, the proposed multirate DSP model finds $X_l(n)$, $\forall\ l \in \mathbb{F}$, using $\{2R\}$ real multiplications and $\{M+2R\}$ real additions, compared to $\{2M\}$ real multiplications and $\{2(M-1)\}$ real additions for the direct DFT implementation as summarized in Table 1. Besides these quantitative results, the implementation cost can be further reduced knowing that typically $R \ll M$ and that $\{M\}$ out of the $\{M+2R\}$ additions required by the multirate analysis are actually simple increments.

Model	Multiplications	Additions
Multirate	$2R$	$M+2R$
Direct DFT	$2M$	$2(M-1)$

Table 1: Real computations required by the multirate model versus the direct DFT implementation $\forall\ l \in \mathbb{F}$.

6. CHOOSING THE DIGITAL FILTERS

Consider again the proposed model of Figure 3. For simplicity, we confine our discussion to the case where $H_r(z) = H(z)$, $\forall\ r$. The exposition of sections 3 and 4 focussed on the special case where $H(z)$ is a rectangular window. Figure 7(a) shows the magnitude response $|F(z)| = |C(z)H(z^5)|$ when $H(z)$ is the standard rectangular window and $M=500$. Although $F(z)$ is a complex bandpass filter with center frequency at $\omega = 2\pi/5$, it displays a poor magnitude response. A sharper response can be obtained by using a more general FIR filter $H(z)$. General FIR windows $H(z) = \alpha_0 + \alpha_1 z^{-1} + \ldots + \alpha_{(\frac{M}{R}-1)}\ z^{-(\frac{M}{R}-1)}$ can be therefore utilized to replace the rectangular case in the model of Figure 3. Table 2 compares the noise rejection performance of four standard windows. Obviously, the Blackman window enjoys a highly attenuated first side-lobe which is the main parameter

affecting the noise attenuation. Figure 7(b) shows the magnitude response $|F(z)| = |C(z)H(z^5)|$ when $H(z)$ is a Blackman window and $M = 500$. The center frequency is at $\omega = 2\pi/5$ but all harmonics are below -40 dB and the first side-lobe goes down to almost -60 dB. Hence, the *filtered* model will accurately *pick* the required frequency component. Finally, a potential boost in performance can be obtained by *biologically optimizing* $H(z)$ or more generally, $H_r(z)\ \forall\ r$.

FIR window	A_1/A_0	$\Delta\omega_m$	$20\log_{10}\delta$
Rectangular	-13	$4\pi/(M+1)$	-21
Bartlett	-25	$8\pi/M$	-25
Hamming	-41	$8\pi/M$	-53
Blackman	-57	$12\pi/M$	-74

Table 2: Comparison of common windows: peak side-lobe amplitude A_1/A_0 in dB, approximate main lobe width $\Delta\omega_m$, and peak approximation error δ in dB.

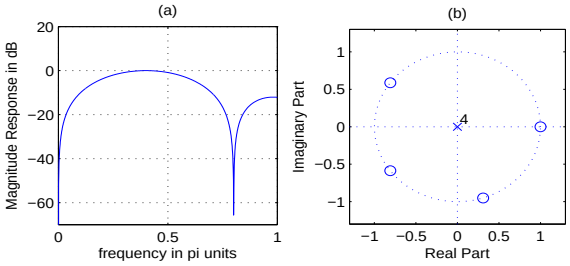

Figure 5: Complex filter $C(z)$: (a) Magnitude response $|C(e^{j\omega})|$, (b) zero-pole plot, for $R = 5$.

7. REFERENCES

[1] B. D. Silverman and R. Linsker, "A measure of DNA periodicity," *Journal of Theoretical Biology*, vol. 118, no. 3, pp. 295–300, February 1986.

[2] D. Holste et al., "Repeats and correlations in human DNA sequences," *Physical Review*, vol. E 67, no. 06913, 2003.

[3] V. R. Chechetkin and A. Y. Turygin, "Search of hidden periodicities in DNA sequences," *Journal of Theoretical Biology*, August 1995.

[4] S. Tiwari et al., "Prediction of probable genes by Fourier analysis of genomic sequences," *CABIOS*, vol. 13, pp. 263–270, June 1997.

[5] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, July 2001.

[6] A. Rushdi and J. Tuqan, "Gene identification using the Z-curve representation," *in the proceedings of the* 31^{st} *IEEE ICASSP conference*, pp. II.1024–II.1027, May 2006.

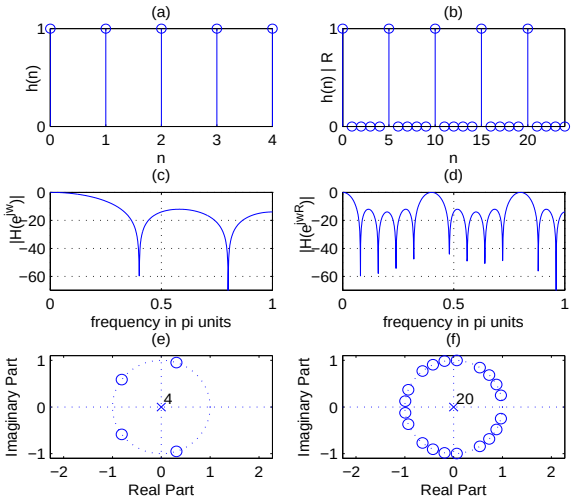

Figure 6: (a),(b) Impulse responses $h(n), h(n)|_{\uparrow R}$, (c),(d) magnitude responses $|H(e^{j\omega})|$, $|H(e^{j\omega R})|$, and (e),(f) zero-pole plots of the real LP filters $H(z)$, $H(z^R)$ respectively, for $M = 25$, $R = 5$.

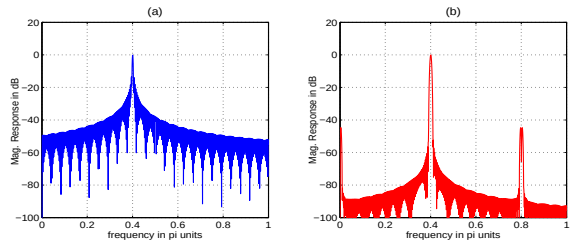

Figure 7: $|F(z)|$ when $H(z)$ is (a) rectangular, (b) Blackman window, for $R = 5$ and $M = 500$.

[7] J. Tuqan and A. Rushdi, "A DSP perspective to the period-3 detection problem," *in the proceedings of the* 4^{th} *IEEE GENSIPS conference*, pp. 53–54, May 2006.

[8] A. Rushdi and J. Tuqan, "The filtered spectral rotation measure," *in proceedings of the* 40^{th} *IEEE Asilomar Conference on Signals, Systems, and Computers*, October 2006.

[9] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions," *Genome Research*, July 2003.

[10] E. N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences," *Physica A*, vol. 249, pp. 511–516, June 1998.

[11] J. A. Berger, S. Mitra, and J. Astola, "Power spectrum analysis for DNA sequences," in *Proceedings of the Int. Sym. on Signal Processing and its App.*, July 2003.